13281 U.S.PTO
031004

17302 U.S.PTO
10/795976
031004

Docket No.: 3351-028A

IMAGE PAGE SEARCH FOR ARBITRARY TEXTUAL INFORMATION

Related Application

[001] This application is a continuation-in-part application of Serial No. 09/126,818 filed July 31, 1998.

Field of the Invention

[002] The present invention relates generally to imaged document searching, and more specifically, to an imaged document search that uses optical correlation methods and techniques to search for arbitrary textual information contained in imaged documents. The result is a significant advance in high-speed search for textual information within imaged documents.

Background of the Invention

[003] Numerous organizations must review and extract information from large repositories of imaged documents. Document images may contain information in a variety of languages and can be printed or handwritten. These document images are not able to be directly searched using typical information retrieval techniques because the contents are represented as pixel collections instead of computer language characters.

[004] Organizations attempting to exploit information from image document pages are the very last link in a complex chain of circumstances that effect the quality of the pixel collections that the organization is attempting to manipulate.

[005] First, the page producer selects a typeface and size (e.g. Times 14) for use in imaging the textual information, producing a particular visual appearance. Next, the page producer selects a particular hardware device (e.g. Epson 940 inkjet printer) to produce the paper copy; different printers will affect the visual representation significantly. After production, the page may be subjected to a variety of processes that may alter the visual

representation of the page. The page may be copied using copier devices that introduce distortions or other visual artifacts. The page may be subjected to environmental insults such as being crumpled or obscured with dirt or liquid. Finally, when the page is to be scanned into the database that the system will be using for search operations, the visual representation of the page image will be influenced by the quality and characteristics of the scanner used as well as the quality of the scanning technique employed.

[006] Most approaches to the problem of searching imaged documents start with an initial step of converting written content from an image format to electronic text. Traditional solutions are based on optical character recognition (OCR) techniques, which have numerous problems. First, as discussed above, document images may be in less than ideal condition. Distortion, rotation, duplication artifacts, or transmission noise may be present and can preclude effective OCR processing. Second, the OCR conversion process can be too slow to cope with required document processing speeds. Third, normal error rates in OCR conversion have a significant negative impact on downstream use of the textual information. Fourth, there are many languages for which there are no OCR conversion engines at all or no engines of acceptable quality.

[007] Because of these problems with existing practices, an approach was needed to search for arbitrary written information contained in imaged documents directly eliminating the OCR process. This approach of the present invention is called optical word recognition (OWR). The present invention advantageously uses techniques to search for arbitrary textual information contained in imaged documents. The result is a significant advance in high-speed search for textual information within imaged documents. The present invention can be used, for example, in language identification, signature identification and signature detection. It is especially useful in searching for the images in large databases.

Summary of the Invention

[008] It is therefore an object of the present invention to provide a method of automatically identifying a pattern on a page including synthetically generating textual patterns as signal templates and compensating, if necessary, for visual differences between the synthetically generated textual patterns and images compared against the

synthetically generated images and compared to compensated images against images in a database.

[009] Another object of the present invention is achieved by a computer software product configured to automatically identify a pattern on a page that includes the computer software product including a medium readable by a processor. The medium has stored a first sequence of instructions, when executed by the processor, causes the processor to synthetically generate textual patterns as signal templates. A second sequence of instructions when executed by the processor causes the processor to compensate, if necessary, for visual differences between the synthetically generated textual patterns and images being compared against the synthetically generated images. A third set of instructions, when executed by the processor, causes the processor to compare compensated images against images in a database.

[010] These and other objects of the present invention are achieved by an optical apparatus configured to automatically identify a pattern on a page. The optical apparatus includes a generating unit for synthetically generating textual patterns as signal templates. The optical apparatus has a compensating unit for compensating, if necessary, for visual differences between the synthetically generated textual patterns and images being compared against the synthetically generated images. The optical apparatus has a comparing unit for comparing compensated images against images in a database.

[011] The foregoing and other objects of the present invention are achieved by a computer-readable medium configured to automatically identify a pattern on page. The computer-readable medium has stored a plurality of sequences of instructions, the plurality of sequences of instructions which, when executed by a processor, causes the processor to perform. The computer-readable medium synthetically generates textual patterns as signal templates. The computer-readable medium compensates, if necessary, for visual differences between the synthetically generated textual patterns and images being compared against the synthetically generated images. The computer-readable medium compares compensated images against images in a database.

[012] The foregoing and other objects of the present invention are achieved by a computer system for automatically identifying a pattern on a page. The computer system is comprised of a processor and a memory coupled to the processor. The memory has

stored sequences of instructions which, when executed by the processor, synthetically generates textual patterns as signal templates. The computer system compensates, if necessary, for visual differences between the synthetically generated textual patterns and images being compared against the synthetically generated images. The computer system compares compensated images against images in a database.

[013] Still other objects and advantages of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein the preferred embodiments of the invention are shown and described, simply by way of illustration of the best mode contemplated by carrying out the invention. As will be realized, the invention is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the invention. Accordingly, the drawings and description thereof are to be regarded as illustrative in nature, and not as restrictive.

Brief Description of the Drawings

[014] The present invention is illustrated by way of example, and not by limitation, in the figures of the accompanying drawings, wherein elements having the same reference numeral designations represent like elements throughout and wherein:

Figure 1 is a flow chart of a method of synthetically generating textual patterns according to the present invention; and

Figure 2 is a flow chart of an image comparison technique according to the present invention.

Detailed Description of the Invention

[015] A method and apparatus of using optical word recognition is described. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide for a thorough understanding of the present invention. It will be apparently, however, that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

[016] As explained in Application Serial No. 09/126,818 filed July 31, 1998, entitled "Imaged Document Optical Correlation and Conversion System" (hereinafter called "the '818 application"), signal templates are generated by selecting a portion of the image. As discussed in this application, the signal templates used to search documents are synthetically generated, allowing searches to be independent of the: a) font type and font size used in a given page being searched; b) characteristics of the printer used to produce the page; c) page image quality due to degradation of the original, d) page image quality due to poor quality scanning, e) operator's ability to find an exemplar of the desired search term in an existing page. These signal templates are compared using a type of comparison performed using optical or digital techniques (as discussed below).

[017] This difference removes the transition costs (in terms of processing time) from digital representation to optical representation for the correlation. The present invention also allows the search process to be performed using inexpensive general purpose computing systems rather than expensive optical correlation equipment. The present invention further allows individual calculations in the search operations to be distributed and performed in parallel using general purpose cluster computing systems.

[018] In the present invention, two pages identified as duplicates are assessed for quality and the lower quality page deleted, as opposed to the '818 application, where duplicate documents are simply deleted.

[019] The '818 application refers to the selection of a pattern from a document page for use as a signal template. This is an ingenious, but flawed way to eliminate very complex parts of the search process. By using a pattern selected from a pre-existing document, one does not have to compensate for the factors discussed in the page production paragraph in the Background section. Unfortunately, this places severe limits on the ability of the invention discussed in the '818 application to successfully search for patterns that differ in any way from the pattern selected from the pre-existing document. Obviously, many pages being searched will differ in how they were composed, printed, handled and scanned.

[020] The present invention takes an alternative path, synthetically generating textual patterns for use as signal templates in order to produce a compensation process necessary to accommodate the crucial visual representation differences between different font

typefaces, different font sizes, as well as distortions introduced in the subsequent printing, handling and/or scanning of the page.

[021] The flow of this target generation and search process is as depicted in Figure 1 starting at step 5. The process starts at step 10 with a search word from the user specified using a numeric representation of the characters in the search word using Unicode (allowing search terms in any written language). At step 20, a database of pages is used to search for the word. Step 10 is performed for each page searched.

[022] Advantageously, the synthetic generation process makes use of information (page metadata) automatically developed for the page at the time it was initially added to the page database. At step 30, the page metadata is used to decide which font typefaces and font size combinations are necessary to represent the fonts known to be present on the page. Thus, one search word may produce several search patterns; for example, the word in Times/12, Times/10 and Arial/18. Step 20 is performed for each page searched. At step 40, page distortion information is identified for each page and is provided to step 70 as discussed below. At step 50, the needed versions of search words using page fonts are rendered using page metadata 30 and search words from the user 10. At step 60, pattern collections are searched using rendered versions from step 50. At step 70, the page metadata is also used to compensate for distortions (if any was found) of the page through the printing, handling and scanning process. The compensations may include small enlargements or reductions in search pattern size, visual distortions such as rounding off of fine details, or other necessary modifications. Thus, each of the search patterns developed in the previous processing step 60 may be modified (or modified versions added to the pattern collection) producing a final collection of search patterns at step 80. At step 100, additional revisions of search patterns with distortion compensations are rendered and the results are produced at step 110. Then the entire above process (steps 10-100) is repeated for the next page to be searched.

[023] There are numerous variations on this basic process, all involving optimizations such as generating all font type/size combinations of the search term in advance, but regardless of these techniques, the basic process (if not the sequential order) will remain essentially the same.

[024] With respect to comparison steps 70 and 100 in Figure 1, other correlation methods, such as digital techniques, can be used than the method described in the '818 application. A more general form of image comparison can be performed using techniques other than optically-based correlation. This comparison technique could be considered to achieve ends equivalent to optical correlation but through different means.

[025] The core of the comparison approach according to the present invention is the following algorithm as illustrated in Figure 2. The process is started at step 200. At step 210, the document page to be searched is acquired in an image format. At step 220, the page image is reduced in resolution, inverted and mirrored. A two-dimensional Fast Fourier Transform (FFT) moving this representation from the spatial to frequency domain is performed at step 230. The search target image (a document, word or other image) is reduced in resolution at step 240 and then an FFT is performed at step 245. The FFT images of the document and the target are multiplied at step 250 to produce a correlation plane and an inverse FFT is performed at step 260 to take it from the frequency spectrum to the spatial specification to produce a similarity matrix for the search pattern locations within the document image. A threshold is then applied to the matrix, and the locations of matches above the threshold are extracted at step 290.

[026] Next, to increase accuracy, these candidate matches are processed through additional comparison processes, but instead of using the entire page image as a comparison element, only the segment of the page image corresponding to the candidate search result is used. This filtering process may use multiple algorithms at step 290. In the current invention, the key discriminating filter used is a spatial domain comparison technique called void space filtering. In this technique, an inverse connected element analysis technique discovers large blocks of white space surrounding the image to produce a "fingerprint" of the white space in an image. Such fingerprints can be created and compared very quickly, and are thus well suited to this application.

[027] An addition to avoid space filtering, second stage comparison algorithms similar to the FFT based technique above, differing in that resolution reduction is not performed, can also be used. After all filters have been applied to the comparisons, successful matches are extracted as the search results at step 310. The process ends at step 320.

[028] In the current invention, two pages identified as duplicates are assessed for quality and the lower quality page deleted. Quality assessment is accomplished by performing a connected element analysis to identify "speckle" (indicating degradation in the handling or scanning processes) as well as blocks of solid color (indicating portions of fully saturated text such as would appear in redact or obscured text). Other methods for quality assessment are possible.

[029] It should now be apparent that a method has been described in which images can be searched from a database and the optical correlation has been eliminated. Advantageously, the inventive method compensates for distortions caused by printing, handling or a scanning process.

[030] It is readily seen by one of ordinary skill in the art that the present invention fulfills all of the objects set forth above. After reading the foregoing specification, one of ordinary skill will be able to affect various changes, substitutions of equivalents and various other aspects of the invention as broadly disclosed herein. It is therefore intended that the protection granted hereon be limited by the definition contained in the appended claims and equivalents thereof.